

AD-A085 896

ILLINOIS UNIV AT URBANA-CHAMPAIGN COORDINATED SCIENCE LAB F/G 5/7
GENERATING AND UNDERSTANDING SCENE DESCRIPTIONS.(U)

MAR 80 D L WALTZ

N00014-75-C-0612

UNCLASSIFIED

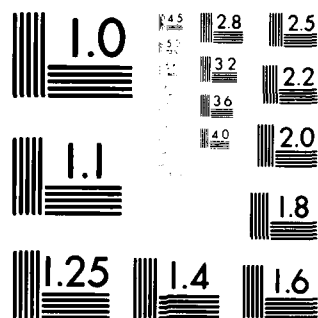
WP-24

NL

(U)
DATE
FILMED



END
DATE
FILMED
*8-80
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

ADA 085896

LEVEL II

(12)

GENERATING AND UNDERSTANDING SCENE DESCRIPTIONS.

9 Working Paper 24

10 David L. Waltz

Coordinated Science Laboratory and
Electrical Engineering Department
University of Illinois at Urbana/Champaign
Urbana, IL 61801

11 March 1980

Abstract

DTIC
ELECTE
S JUN 24 1980 D
E

This paper explores design issues for a system which has both vision and language, in particular, a system which addresses both the problem of selecting appropriate words and sentences to describe a particular perceptual event, and the related problem of making appropriate inferences about a natural language description of a perceptual event. It argues that perception is basically a description-building process, and that the understanding of scene descriptions is ultimately based on our ability to first use scene descriptions to drive processes of 'picture-building', and then to drive processes of 'event-simulation' which cause the "pictures" we build to mimic the dynamics of the world.

This work was supported by the Office of Naval Research under Contract N00014-75-C-0612.

This paper will appear in Joshi, Sag and Webber (eds.) Elements of Discourse Understanding, Cambridge University Press, 1980.


DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

-1-

80 5 5 074

DDC FILE COPY.

| | |
|-------------------------------------|---|
| Accession For | |
| NTIS GRA&I |  |
| DDC TAB | |
| Unannounced | |
| Justification <i>Letter in File</i> | |
| By | |

1. Introduction

"A semantic theory having no contact with the world, a mere translation of one set of words into another, is a ladder without rungs."

-Miller and Johnson-Laird (1976)

The research reported here is part of a larger project whose goal is to link language and perception, and in this effort to provide a deeper understanding of what it means to understand. Computer vision and language researchers have to date had little to say to each other. However, without a connection to the real world via perception, it is difficult to say in what sense a natural language system could be said to understand descriptions of the physical world. If an entity (human or computer) understands scene descriptions, it should be able to make predictions about likely futures; it should be able to judge certain scene descriptions to be implausible; it should be able to point to items in a scene, given a description of the scene; and it should be able to say whether or not a description corresponds to a given scene. This requires that it have a vision system which can generate scene representations that can be compared with scene representations generated from natural language descriptions. This in turn requires that the entity be able to pay attention to what is important in context, must be able to note events and cause-effect relationships over time, and in general be able to find patterns in (or impose organization on) visual data.

This paper explores design issues for a system which has both vision and language, in particular, a system which addresses both the problem of selecting appropriate words and sentences to describe a particular perceptual event, and the related problem of making appropriate inferences about a natural language description of a perceptual event. I first consider (in

section 2) the problem of understanding scene descriptions, and concentrate on the problems of judging the plausibility of described scenes, and making inferences about the causes and effects of described events. To account for our ability to handle these problems, I introduce "event simulations", procedures which model events. I then in section 3 consider scene description generation, providing some connections between processes of perception and processes of description generation. In particular, I argue that perceptual processes also activate event simulations in the perceiver; and that scene descriptions are often stated in terms of the names of event simulations because the event simulations can associate and refer compactly to a number of scene items and relationships, and also because a major goal in scene description is to help a hearer to set up event simulations similar to those of the speaker/perceiver. Finally, in section 4 I give a brief historical perspective on related research, and speculate on how my research might be extended to make possible language/vision systems which could learn from experience.

1.1 The domain of interest

Our earliest language is primarily concerned with describing the perceived world. Both for infants and adults the outside world is always filtered through and confounded with an internal world of interpretation: The outside world is alternatively interesting, boring, peaceful, threatening, pleasurable, and painful; items in the outside world are in varying degrees similar to other items seen or remembered. Nonetheless, it seems to us that we learn early to factor out a neutral outside world from our internal world, so that we can produce descriptions of the outside world which are intelligible to others, and so that we can understand others' descriptions of objects, events, and relationships between them. It is the perception of this "neutral" outside world as expressed in

language, and the understanding of language describing this outside world which are the central concerns of this paper.(1)

I will explore the following questions:

(1) What processes allow me to understand and describe scenes I encounter? How do I decide what to include in a description?

(2) If I hear a sentence (e.g. "The car hit the boy.") what inferences can I make about the perceptual world of the speaker?

(3) Under what perceptual circumstances is it appropriate to use particular words and sentences, i.e. what things could I be looking at to appropriately report a sentence like "The car hit the boy."?

(4) What knowledge and procedures would be necessary for a program to produce and understand this sort of language?

In concentrating on perceptual circumstances, I am explicitly deemphasizing important questions about why it would be appropriate to utter a description of the world; for now I would like (as much as possible) to factor out issues of intent in making utterances, and concentrate on the issues involved in "accurate reporting" only. As we will see, it is difficult to separate these issues.

1.2 Components

What would a total system be like? I suggest that it would at least have to include components(2) for handling each of the following subproblems:

(1) I will later argue that neutrality is impossible, since we always act because of goals, and these goals affect every aspect of cognition from attention to interpretation.

(2) Alternatively one could argue for a totally integrated system with no distinct component boundaries. Very little of my discussion depends on the assumed decomposability of the problem.

Syntax and semantics We need the ability to parse natural language about the physical world, the ability to choose appropriate senses of words (though this may often involve only the process of choosing the "physical interpretation" of a word from the lexicon), and the ability to organize the information into an appropriate "deep structure" (first order predicate calculus or some other adequate form). I will not say much directly about these components, although some of the issues and examples I discuss have a bearing on the design of these components.

Pragmatics As mentioned above, for now I wish as much as possible to factor out consideration of the motives and goals for uttering scene descriptions. However, under the section on "appropriateness" are included various factors and criteria for deciding which words to use in describing a scene, and these have a distinctly pragmatic flavor.

Vision I assume the possibility of designing a vision system substantially different from those that have been constructed to date. The chief difference is that the system I want does not perform a total analysis of each static "frame" of a scene, but instead (1) works on a constantly moving scene image, (2) alternately attends to an "item" -- some small portion of the overall scene -- and shifts attention to another item, by moving through some trajectory. The vision system needs criteria for deciding where to look next; such decisions are to be based partially on what is most "interesting" in the periphery of the current visual field, and partially on the current goals, tasks, and hypotheses of the system. We are a rather long way from constructing a vision system of this sort; while I have some ideas on the design of such a system, I will not be concerned here with the details of how it might be done.

World knowledge - epistemology The major emphasis in this paper is on the representation and use of world knowledge. I am particularly interested in how a scene description is built up and organized, and in how comparable structures could be constructed from language inputs.

1.3 Fitting the pieces together

I believe that it is important to consider fitting together all the components mentioned in the previous section. First, it will be practically impossible to actually build a system of the sort I am describing unless the design is considered in toto. But more importantly, the kinds of solutions offered to each of the areas (syntax, semantics, pragmatics, epistemology, vision) may be totally incompatible if each area is only considered independently; in addition there is a danger that important problems may "fall between the cracks" separating the areas, and never be considered at all.

2. Understanding scene descriptions

In this section I argue that scene descriptions are most naturally treated by representations which are, at least in part, only awkwardly viewed as propositional; such representations include coordinate systems, trajectories, and event-simulating mechanisms.

Compare the following sentences:

(S1) My dog bit the mailman.

(S2) My dog bit the mailman's leg.

(S3) My dachshund bit the mailman's ear.

(S1) and (S2) do not seem to require visualization for understanding (although they may evoke mental images for some people). SCRIPT-like formalisms (Schank & Abelson 1977) seem at least on the surface to be extendable for expressing such

sentences internally, although a BITING script would involve a dog's goals rather than the human goals found in most scripts. However, I believe that many (possibly most) people would note that something is wrong or peculiar about sentence (S3). Even if people did not catch the fact that (S3) is peculiar, once the sentence is pointed out, it does seem to "require an explanation" because it is not possible to figure out with certainty how the dachshund was able to bite the mailman's ear without further information. (Possible explanations: the mailman fell while trying to get away from the dog; the mailman was kneeling or squatting to pick up a dropped item or to pet the dog.) How can we judge that an explanation is needed? What mechanisms could we use to understand an explanation, and what is the relationship between an explanation and the thing explained?

One possible answer to the first question is that we have saved all biting sites ever encountered as possible fillers of a slot in a biting script. When we encounter a bitten ear, we note that it is unusual or never before encountered. This possibility seems to be ruled out, at least as a full explanation, because

(S4) My doberman bit the mailman's ear.

does not seem peculiar in the sense that (S3) does. It seems to be that we really make a judgment that a dachshund could not reach a person's ear ordinarily. The possibility that we store ahead of time information of the form: (<dog type>, <part of the body>) for each type of dog and reachable part of the body seems too remote to me to consider seriously.

2.1 Event Simulation

If we don't prestore such lists of possibilities, the only alternative seems to be to compute them when needed via some mechanisms. But if we believe (or can show via psychological testing) that people readily catch physically unlikely sentences

like (S3), then it may be that we compute physical plausibility always (unconsciously) or that event simulation mechanisms are invoked by higher level exception-trapping mechanisms. In any case, it seems to me that event-simulating mechanisms are necessary for full understanding of language about the sensory world. (1) I say this for the following reasons:

(1) We are able to make plausibility judgements about descriptions. (2)

(2) Something like event simulation seems necessary for the resolution of anaphoric reference. Event simulations could help both in setting up expectations, and in attempting to set up plausible "pictures" for the purpose of comparing various pronoun reference candidates. (3)

(3) Event simulations may allow us to circumvent short term memory limitations. It seems to be possible to remember entire pictures as though they were rather like single chunks. To the extent that items and events can be pictured, it may be possible to enlist perceptual apparatus in reasoning, and thereby achieve greater power and efficiency. Along similar lines, experts in memorization have long used visual imagery to aid retrieval of items ([Luria 1968], [Bower 1970]).

(4) There are also likely to be ties between this work and current research on "mental imagery" [Kosslyn et al 1979]. While I see no compelling argument that event simulation must give rise to mental images, the existence of mental imagery seems to me to require the existence of some mechanisms like those of event simulation.

(1) I am aware of the dangers in posulating the existence of mechanisms on the basis of our behavior in exceptional cases (e.g. understanding sentences such as (S3)). It might be possible to get more convincing evidence through psychological testing. If event simulation shares resources with the visual system, then it should be possible to show interference between the understanding of scene description and perceptual tasks.

(2) Plausibility judgement may perhaps be usefully viewed as a

Let us return to the dachshund example, and be more specific about what event simulation would involve in this case. The "event simulation" would (1) "create" a mailman and dachshund in default positions (both standing) on level ground outdoors with no special props other than the mailman's uniform and mailbag; (2) test to see if the dachshund can reach the mailman's ear with its mouth directly (no); (3) see if the dog can stretch or jump high enough to reach it (no); (4) see if the mailman would ordinarily get into positions where the dog could reach the ear (no); (5) judge that the mailman could not be bitten as stated unless default states and movement ranges are relaxed. Since there is no clearly preferred way to relax the defaults, more information should have been included in the description, according to the criteria for scene descriptions listed below (section 3). Speakers should realize the need for more information because they should run event simulations on their own output; if for some reason a speaker has not kept track of the picture suggested to the hearer, (s)he should be able to construct the picture rapidly if the hearer hasn't understood the description.

Some other examples of sentences which fail event simulation verification are listed below. Some of these (e.g. S5 and S6) are probably answered via recourse to "world knowledge" rather than actual event simulation.

(S5) I ate 50 eggs for breakfast yesterday.

(S6) My cat killed an elephant.

(S7) I divided the birthday cake 1000 ways.

(S8) The mouse ran across the hood of my car and dented it.

(S9) The rock floated toward the shore.

(S10) A 747 flew so low that it knocked the top of my chimney off.

deeper analog of grammatical and semantic judgements.

(3) I am grateful to Candy Sidner for pointing this out to me.

- (S11) The small urn of oil burned for eight days.
(S12) We managed to stuff 20 people in a phone booth.
(S13) The tree grew four feet overnight.
(S14) My car hit the telephone pole at 55 miles per hour, but wasn't damaged.
(S15) Butterflies surrounded us while we skated on the pond.
(S16) I dropped a rock from the window and ran downstairs in catch it.
(S17) The hot water I spilled melted the stove.

Many of these examples are reminiscent of the Guinness Book of World Records or of miracles from the Bible. In each case our simulation of the described event is at least difficult to believe, in all cases contrary to ordinary experience.

2.2 Proposed mechanisms for event simulation and scene description generation

I have been working on a sketch of a design for a system to provide event simulations, given sentences as input. The system depends on (1) a large taxonomy of event types, with structural inheritance links so that events may be treated with a wide range of precision or generality; (2) time sequencing information for events, so that events may be ultimately broken down (if necessary) into very low-level "primitives" or aggregated into larger event units, and so that a program can predict effects of events and can infer likely causes of events. This taxonomy is built so that it can function as a kind of decision tree for perceptual processes; however it has words attached to event types in the taxonomy in such a way that it can be used both to simulate events during text understanding, and to generate "appropriate descriptions" and expectations for scenes and events if driven by a perceptual system.

One major piece of a system to run event simulations is Rieger's CSA ("commonsense algorithm") system [Rieger 1975].

CSAs break the world into STATES, STATECHANGES, ACTIONS, TENDENCIES, and GOALS; these can be interconnected with about twenty different causing and enabling relationships (e.g. continuous causation, as pressure causing flow; gated enablement, as in an open valve enabling flow; one-shot causation, as in pushing the flush handle on a toilet; etc.). While a full discussion is beyond the scope of this paper, CSAs have been used to model some physical systems (electronic circuits, a toilet, the process of combustion). It will at least be necessary to augment CSAs via (1) the addition of time and quantities in general -- CSAs are now primarily qualitative; (2) the addition of spacial information -- coordinate systems, dimensions, etc.

3. Deciding what to include in a description

In this section, I want to consider the problem of choosing appropriate words to describe events in a scene, concentrating particularly on verb choice. Much of what I say here may apply as well to the choice of other words and to the form of sentences (e.g, which item is chosen as syntactic subject, which material is put in relative clauses and which in main clauses, etc.). Basically I argue in the following sections that:

(1) scene descriptions should include items noticed or inferred, and subsequently judged to be important;

(2) scene descriptions should satisfy certain criteria of "appropriateness" which insure that a hearer will be able to build a plausible, coherent internal representation of the description;

(3) scene descriptions depend on available vocabulary and language production procedures;

(4) scene descriptions always serve some goals of the speaker.

This last point deserves added emphasis. I contend that there is no such thing as a purely objective scene description. What we call "objectivity" can more accurately be described as realizing the goal of reporting on all the objects and events in a scene, and at the same time concealing one's opinions and evaluations of the scene. Even for simple scenes it is impossible to cover all the things that could be said about the objects and events in it; n objects in a scene can be grouped in $n!$ ways, and each of the groups may be describable in many ways (e.g. by focusing on different elements in a group), and the groups may be considered in any order as can choices of focus. Furthermore, the plausible origins of the scene, the expected future of the scene, the reasons that the speaker is generating the description in the first place, and the reasons why the speaker has chosen the particular order of description all add open-ended possibilities for scene description that cannot be neatly separated from the scene per se.

3.1 Attention and salience

Items can only be part of a description if they have been noticed, or if they can be inferred from what has been noticed. What is noticed is in turn a complex function of one's goals and the context of the scene. This section lists a number of factors that affect what we attend to, infer, and remember -- the "raw material" of descriptions.

(1) External factors Motion, contrast, size, color, complexity, symmetry, asymmetry, density of interesting features, plus many other scene characteristics can attract attention or camouflage items. We sometimes say of striking items "You can't miss it" (though we often do).

(2) Internal factors At the same time goals, desires, habituation, familiarity, novelty, and other internal factors affect attention. (In the words of a proverb: "A thief looks at a saint and sees pockets.") Some items probably seem important to us because they activate mechanisms that have been evolved or conditioned to decide whether a scene contains items that are valuable to us or that threaten us. (As Bill Woods has put it, we constantly ask the questions: "Can I eat it? Can it eat me?")

However, most internal factors are goal-dependent. For example consider a particular outdoor scene; we notice very different items and relationships depending on our current goals. If I am looking for a lost wallet I will attend to places in the scene where I think I have been, objects that might be the wallet, and objects that might obscure the wallet. The processing will be very different, however, if I am looking for a good place to have a picnic, or trying to figure out where I am, or hunting for firewood, or playing hide and seek with my children (and then somewhat different depending upon whether I am the hider or the seeker).

(3) Vantage point One's position with respect to an event affects the relationships one sees between objects in the visual field; the inability to see parts of a scene can lead to hedged descriptions as in "I think that John hit Mary first." However, at least to a degree, we can "see" events independent of viewpoint. For example, we see (and describe) "two cars approaching one another" and not "one car moving left-to-right and another car moving right-to-left". We can also include point of view in our descriptions by using orienting phrases whose meaning can be shared, e.g. toward the north, away from the house, on his left, etc.

3.2 Appropriateness

To be "appropriate", descriptions of a scene should meet the following interrelated criteria. These criteria have close ties with Grice's principles of cooperative conversation [Grice 1975]; however, I arrived at these criteria by looking at a number of examples of descriptions, and generalizing my observations.

(A) Descriptions should include all items attended to and subsequently judged to be important,

(B) Descriptions should be as economical in the use of clauses and words as possible, (1)

(C) Descriptions should use words and structures whose implications and attached default assumptions are actually true of the scene. If the words one wants to use invite inaccurate inferences, the inaccurate inferences must be explicitly ruled out or modified.

It is on criteria (B) and (C) that I would like to concentrate. By (B), we would prefer description (S18) to description (S19):

(S18) He knocked the glass onto the floor.

(S19) He hit the glass and knocked it onto the floor.

This because "He hit the glass" can be inferred from (S18). Similarly, criterion (B) favors (S20) over (S21):

(S20) Two cars collided head-on.

(1) Obviously with children or people who are unlikely to understand certain words (e.g., technical terms), one uses non-optimum descriptions. I used to tell my children "Our ride will be as long as Captain Kangaroo's show" because they didn't understand directly how long an hour is.

(S21) One car was moving on the road and another car was moving on the road in the opposite direction, and the two cars hit each other.

This is because (S20) is a paraphrase of (S21) which uses many fewer words and clauses.

Criterion (C) is a more subtle. Suppose that an automobile just grazed a boy or hit just his finger as it went by. Strictly speaking, it would be possible in either case to say:

(S22) The car hit the boy.

However this description is misleading, since the default assumptions one would make if (S22) were heard out of context are that a major part of the boy's body was struck. (More precisely, the boy's center of mass was probably situated within the volume defined by projecting the car's frontal cross-section forward along the car's trajectory.)

Example sentence (S3)

(S3) My dachshund bit the mailman's ear.

is also inappropriate if uttered out of context, since in order to understand (S3) one must invoke explanations that violate default assumptions, e.g. that the mailman is standing, the dachshund is of ordinary size and on the ground, etc. One cannot complete the internal representation of the scene corresponding to (S3) with any confidence that the completion represents the actual situation in the world.

Understatement and overstatement are also violations of criterion (C):

(S23) My car was damaged in an accident.

would be inappropriate if the car were a total loss, since "damaged" entails the possibility of repair, unless specifically excluded as in the phrase "damaged beyond repair."

3.3 Expressibility

Items can only appear in the same clause if they all belong to a single perceptual pattern, (e.g. event or cause-effect relationship) and if words are available to predicate this patterned relationship between items. Obviously one can simply list all the items present in a scene whether related or not -- here the perceptual pattern is simply: "present in the same scene". If the items cannot be seen as part of a single event, however, they have to be described separately, using structures such as: "<event-description-1> and meanwhile <event-description-2>". Expressibility may seem to be a rather amorphous factor, but I do want to include some indication that our descriptions are constrained by our ability to see the items in a scene as an instance of a pattern, and are also constrained by our vocabulary (lexical and structural) for referring to such patterns.

3.4 Speech acts

The scene descriptions we generate may be more or less inappropriate or may be modified, because our goals (conscious or unconscious) can affect our choice of items to attend to, and can also change our evaluation of the importance of scene items attended to.

The effects of human values are also evident in both the items attended to and in our judgment of the "appropriateness" of descriptions (see below). For example, in a description of an accident where a car hit a boy, we would expect the consequences to the boy to be described first, even if he were not injured. To describe damage to the car first, unless the damage were particularly unusual, would seem at least

inappropriate, and possibly perverse. Furthermore, one's description of an accident would almost certainly exaggerate its seriousness if the victim were one's child and the driver a stranger, but minimize its seriousness if the victim were a stranger and the driver a friend. One would look for evidence to support the belief that the stranger in each case was responsible for the accident, and one would probably also attend especially to the consequences of the accident for one's friend or relative.

Sometimes context can make ordinarily inappropriate descriptions appropriate. For example, if one were asked as an accident witness about whether or not a car had contacted the boy, it might be appropriate to say (S22) -- "the car hit the boy" -- even if the car had only grazed him.

3.5 Miscellaneous factors

The choice of specific words and sentences may also be influenced by a large number of other factors, including: rhymes or close associations with words used earlier in a discourse; parallel syntactic structures; consistent use of same voice (i.e., active or passive); "Freudian slips," i.e., unintended use of inappropriate words which are related to a speaker's suppressed goals; etc. While such factors are clearly important for an overall theory of cognition and scene description, I will not treat them further here.

4. Why are scene descriptions important?

On the surface it may seem that focusing on the domain of physical events is very restricting. After all, most language is not about the physical world per se, but about the "abstract world" of goals, theories, explanations, stories, reports of combined inner and outside world experience, etc. Nonetheless, I think that the domain of physical events is of central importance; I will attempt to explain why in this section.

4.1 Historical perspective

Most efforts in language processing, both in artificial intelligence and linguistics, have concentrated on transforming strings of words into trees or other structures of words (sometimes surface words, sometimes "primitive" words) or conversely, on producing strings of words from these structures. Most language programs "define" nouns as a conjunction of semantic markers (e.g. animate, human, physical object, and so on). At this time in history, AI vision and natural language researchers have little to say to each other; most of the work which treats language and perception⁽¹⁾ together would I think be considered to lie in the realms of philosophy or psychology.

Moreover, the areas of language processing which could have a bearing on perception have been largely ignored. Very little work has been done on programs to understand language about space, spatial relations, or object descriptions. (But see [Bogges 1978], [Waltz and Bogges 1979], and [Waltz 1979].)

By the same token, current computer vision systems are not able to describe what they "see" in natural language; in fact very few programs can even identify objects within a scene (except for programs which operate in very constrained universes). Furthermore, no vision programs are able to tailor their performance to given questions or tasks (e.g. Where could a lost object be in the current scene? Where am I? How can I find a path to take to get to some object in the scene? etc.). Most vision systems simply produce scene segmentations, labelings or 3-D interpretations of scene portions. Programs are universally capable of only a single mode of operation; there is no analog of an attention mechanism or task-dependent performance. Similarly,

(1) While I intend perception to refer in the human examples to all the senses -- vision, hearing, touch, smell, taste, and kinesthetic -- in the case of computers, only vision has been explored in more than a cursory manner.

no programs I know are able to locate or "point to" scene items, given a natural language description of scene items or their whereabouts.

Piaget [1967] has long argued that an understanding of the sensory-motor world is a critically important first step in developing schemas for concepts in abstract worlds. Jackendoff [1976] and Gruber [1965] have suggested in some detail how sensory-motor schemas might be transferred to abstract worlds via the treatment of abstract items as "metaphorical locations". Other researchers have recognized the importance of these problems. Especially noteworthy is the landmark volume Language and Perception by Miller and Johnson-Laird [1976]. Other work of note in this area can be found in [Minsky 1975], [Woods 1980], [Clark 1973], [Bajcsy and Joshi 1978], [Soloway 1978], [Simmons 1975] and [Novak 1976], [Kuipers 1977], and [Johnson-Laird (this volume)].

4.2 Toward programs which learn from experience

The simulation of cognitive processes has been approached in the past by means which are at the extremes of a spectrum: at one end of the spectrum are "adaptive" approaches which assume that systems begin with a blank slate ("tabula rasa"), and that evolutionary, trial-and-error mechanisms will allow the systems to "learn by experience", much as people do (See for example [Holland and Reitman 1975], [Minsky and Papert 1967]) At the other end of the spectrum are artificial intelligence approaches, which generally attempt to model the knowledge of an adult directly, and ignore problems of learning. It has been argued that "...in order for a program to be capable of learning something it must first be capable of being told it" [McCarthy 1968], so that research has concentrated primarily on problems in the representation of knowledge; learning programs (e.g. [Winston 1970] [Sussman 1973]) are extremely narrow in

their competence, and depend on having a good teacher - hardly like learning from experience. However, there seems to be little hope that adaptive approaches will be even as successful as AI approaches; the search space of possibilities is so large that unless programs begin with enough structure to exhibit interesting behavior to begin with, it is overwhelmingly unlikely that the programs will ever evolve to the point of exhibiting interesting behavior.

What can be done? I suggest that by examining the problem of designing a system which integrates vision, language, and memory we can begin to work from an approach which is somewhere between the two extremes of AI and adaptive modeling. The basic argument is this: people are only able to learn because we begin with a great deal of structure in our perceptual systems (and probably other systems as well); a good starting point for learning would be a system which could generate rich procedural descriptions of events in the physical world, and associate language about the physical world with these descriptions. Learning could then be explored in at least two novel ways: (1) the system could add knowledge of specific events to its memory, and attempt to generalize its experience; and (2) we could use it to investigate the use of rich perceptual schemas for interpreting abstract events.

Of course this is only a starting point. In order to be able to eventually learn about abstract worlds as well, a system must be able to bootstrap itself in some way. Jackendoff [1975] and Gruber [1965] have pointed out evidence that linguistic schemas we develop to describe GO, BE and STAY events in the sensory/motor ("position") world are later transferred via a broad metaphor to describe events in abstract worlds (possession, "identification" and "circumstantial"). Thus we learn to use parallel surface structures for conceptually very different sentences such as:

- (S24a) The dishes stayed in the sink (position).
(S24b) The business stayed in the family (possession).
(S25a) His puppy went home (position).
(S25b) His face went white (identification).
(S26a) She got into her car and went to work (position).
(S26b) She sat down at her desk and went to work (circumstantial).

Along these same lines, there are striking parallels in the structures of Schank's [1975] conceptual dependency diagrams for PTRANS, ATRANS, and MTRANS. Reddy [1977] has described what he calls the "conduit metaphor" for linguistic communication in which we typically speak of ideas and information as though they were objects which could be given or shipped to others who need only to look at the "objects" to understand them. Thus we say "You aren't getting your message across," "She gave me some good ideas," "He kept his thoughts to himself," "Let me give you a piece of advice," etc. (Reddy has compiled a very long list of examples.)

These examples suggest many deep and fascinating questions. It seems clear that the same words and similar syntactic structures can be transferred to describe quite different phenomena. What internal structures (if any) are also transferred in such cases? What perceptual criteria are used to classify events to begin with? Ultimately? How does a child transfer observation to imitation? How are memories of specific events generalized to form event types, and how are the representations of event types related to memories of specific events?

There is also a great deal of prima facie evidence of close ties between perception and the language used by adults to

describe abstract processes such as thinking, learning, and communicating, and to describe abstract fields like economics, diplomacy, and psychology. Witness the wide use of basically perceptual words like: start, stop, attract, repel, divide, separate, join, connect, shatter, scratch, smash, touch, lean, flow, support, hang, sink, slide, scrape, fall, grow, shrink, waver, shake, spread, congeal, dissolve, precipitate, roll, bend, warp, wear, chip, break, tear, etc., etc. While we obviously do not always (or even usually) experience perceptual images when we use or hear such words, I suggest that much of the machinery used during perception is used during the processing of language about space and is also used during the processing of abstract descriptions. I do not find it plausible that words like these have two or more completely different meanings which simply share the same lexical entry.

6. Conclusions

I have examined a number of issues in scene description generation and scene description understanding. This work is part of a larger effort to model via computer programs our understanding of the sensory-motor world. I have argued especially for procedural rather than static representations for knowledge, and have attempted to show the intimate connections between discourse about the sensory-motor world, perceptual processes, and "event simulation" mechanisms. I believe that this research can have important consequences in that, compared with the study of isolated components (e.g. vision, syntax, semantics), the design of a complete vision/language system adds many more constraints on the possible for components. I also believe that a thorough understanding of the sensory-motor world is a necessary precursor to a satisfactory handling of abstract worlds, which are understood via metaphorical reference to the sensory-motor world. In turn, solving these problems is essential if we are ever to be able to model "learning from experience" and to understand understanding.

Acknowledgements

I would like to thank Bill Woods for providing office space, computer time and moral support, Jeff Gibbons, Brad Goodman, and Candy Sidner for helpful comments on an earlier draft, Bonnie Webber for her patience and confidence in me, and Bonnie Waltz for her loving support.

References

Bajcsy, R. and Joshi, A. (1978) The problem in naming shapes: vision-language interface. In D. Waltz (ed.), Theoretical Issues in Natural Language Processing - 2, ACM, New York.

Bogges, L. C. (1978) Computational interpretation of English spatial prepositions. Unpublished Ph.D. dissertation, Computer Science Dept., University of Illinois, Urbana.

Bower, G. H. (1970) Analysis of a mnemonic device. American Scientist 222, 5, 104-112.

Brachman, R. J. (1979) On the epistemological status of semantic networks. In N. Findler (ed.) Associative Networks, Academic, New York, 3-50.

Chafe, W. L. (1979) The flow of thought and the flow of language. T. Givon (ed.) Discourse and Syntax. Academic, New York.

Clark, H.H. (1973) Space, time, semantics and the child. In T.E. Moore (ed.) Cognitive Development and the Acquisition of Language, Academic, New York, 27-63.

Grice, H. P. (1975) Logic and conversation. In Cole and Morgan (eds.) Syntax and Semantics Volume 3: Speech Acts, Academic, New York, 41-58.

Gruber, J. S. (1965) Studies in Lexical Relations. Unpublished Ph.D. dissertation, MIT, Cambridge, MA.

Holland, J. H. and Reitman, J. S. (1977) Cognitive systems based on adaptive algorithms. Tech. Rpt. No. 201, University of Michigan, Computer and Communication Sciences Dept., Ann Arbor, Michigan.

Jackendoff, R. (1975) A system of semantic primitives. In R. Schank and B. Nash-Webber (eds.), Theoretical Issues in Natural Language Processing, ACL, Arlington, VA.

Jackendoff, R. (1976) Toward an explanatory semantic representation. Linguistic Inquiry 7,1, 89-150.

Johnson-Laird, P. (1980) Mental models of meaning. To appear in Joshi, Sag and Webber (eds.) Elements of Discourse Understanding, Cambridge University Press, 1980.

Kuipers, B. J. (1977) Representing knowledge of large-scale space. Tech. Rpt. AI-TR-418, MIT AI Lab, Cambridge, MA.

Luria, A. R. (1968) The Mind of a Mnemonist. Basic Books, New York.

McCarthy, J. (1968) Programs with common sense. In M. Minsky (ed.) Semantic Information Processing, MIT Press, Cambridge, MA.

Miller, G. A. and Johnson-Laird, P. (1976) Language and Perception Harvard University Press, Cambridge, MA.

Minsky, M. L. (1975) A framework for representing knowledge. In P. Winston (ed.) The Psychology of Computer Vision, McGraw-Hill, New York.

Minsky, M. L. and Papert, S. (1967) Perceptrons. MIT Press, Cambridge, MA.

Novak, G. S. (1976) Computer understanding of physics problems stated in natural language. Tech. Rpt. NL-30, Dept. of Computer Science, University of Texas, Austin.

Piaget, J. (1967) Six Psychological Studies. Vintage, New York.

Reddy, M. (1979) In A. Ortony (ed.) Metaphor and Thought, Cambridge University Press, New York.

Rieger, C. (1975) The commonsense algorithm as a basis for computer models of human memory, inference, belief and contextual language comprehension. In R. Schank and B. Nash-Webber (eds.), Theoretical Issues in Natural Language Processing, ACL, Arlington, VA, 180-95.

Schank, R. C. (1975) The primitive ACTs of conceptual dependency. In R. Schank and B. Nash-Webber (eds.), Theoretical Issues in Natural Language Processing, ACL, Arlington, VA.

Schank, R. C. and Abelson, R. P. (1977) Scripts, Plans, Goals, and Understanding. Lawrence Erlbaum, Hillsdale, NJ.

Simmons, R. F. (1975) The Clowns microworld. In R. Schank and B. Nash-Webber (eds.), Theoretical Issues in Natural Language Processing, ACL, Arlington, VA.

Soloway, E. (1978) Learning = interpretation + generalization: a case study in knowledge-directed learning. COINS Tech. Rpt. 78-13, University of Massachusetts, Amherst.

Sussman, G. J. (1973) A computational model of skill acquisition. Tech. Rpt. AI-TR-297, MIT AI Lab, Cambridge, MA.

Waltz, D. L. (1979) Relating images, concepts, and words. Proc. of the NSF Workshop on the Representation of 3-D Objects, University of Pennsylvania, Philadelphia.

Waltz, D. L. and Boggess, L. C. (1979) Visual analog representations for natural language understanding. Proc. of IJCAI-79, Tokyo, Japan.

Wilks, Y. (1975) Primitives and words. In R. Schank and B. Nash-Webber (eds.), Theoretical Issues in Natural Language Processing, ACL, Arlington, VA, 38-41.